Spatial release from masking based on binaural processing for up to six maskers

William A. Yost[a)]
Speech and Hearing Science, Arizona State University, P.O. Box 870102, Tempe, Arizona 85287, USA

(Received 9 September 2016; revised 24 February 2017; accepted 28 February 2017; published online 23 March 2017)

Spatial Release from Masking (SRM) was measured for identification of a female target word spoken in the presence of male masker words. Target words from a single loudspeaker located at midline were presented when two, four, or six masker words were presented either from the same source as the target or from spatially separated masker sources. All masker words were presented from loudspeakers located symmetrically around the centered target source in the front azimuth hemifield. Three masking conditions were employed: speech-in-speech masking (involving both informational and energetic masking), speech-in-noise masking (involving energetic masking), and filtered speech-in-filtered speech masking (involving informational masking). Psychophysical results were summarized as three-point psychometric functions relating proportion of correct word identification to target-to-masker ratio (in decibels) for both the co-located and spatially separated target and masker sources cases. SRM was then calculated by comparing the slopes and intercepts of these functions. SRM decreased as the number of symmetrically placed masker sources increased from two to six. This decrease was independent of the type of masking, with almost no SRM measured for six masker sources. These results suggest that when SRM is dependent primarily on binaural processing, SRM is effectively limited to fewer than six sound sources. © *2017 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4978614]

[GCS] Pages: 2093–2106

I. INTRODUCTION

Since the introduction of the masking level difference (MLD) by Hirsh (1948) and Licklider (1948) in the same issue of the *Journal of the Acoustical Society of America*, and followed by the exploration of the "cocktail party problem" by Cherry (1953), many investigators have studied the detection, discrimination, or recognition of target signals in the presence of masker sounds that were either all co-located with the target or spatially separated in one way or the other from the target. In almost all cases, detection, discrimination, and recognition performance improves when maskers are spatially separated from the target as compared to when they are co-located (see Yost, 1997). This improvement is often referred to as spatial release from masking (SRM; see Litovsky, 2012, for a recent overview). SRM in the azimuth plane is often assumed to be at least partially the result of two related binaural processes that may or may not occur together. First, SRM based on binaural processing may result in the ability to localize the target source at a different location from the masker source(s). As a consequence, it might be easier to attend to one or the other. Second, there might be binaural analyses (see Green and Yost, 1975), such as an Equalization-Cancelation (EC) process, (see Durlach, 1963) that enhance target representations and/or degrade masker representations. This could result in SRM even if neither target nor masker sources can be accurately localized (see Durlach and Colburn, 1978).

The majority of the SRM literature has involved one target and one masker. These experiments usually place the masker source nearer to one ear than the other, with the target often being presented at the midline, so the target sound arrives simultaneously and with equal intensity at the two ears. In these cases the head can "shadow" the masker at one ear relative to the other. As a result there is a larger target-to-masker ratio at the "shadowed ear" as compared to the other ear (sometimes called the "better ear effect"). SRM in such conditions can be the result of this larger target-to-masker ratio at the "shadowed ear," rather than binaural processes associated with localization, binaural analysis or, indeed, other processing. To come closer to isolating the role of binaural processing in SRM, head shadowing can be reduced, if not avoided, by presenting masker sources so they are placed symmetrically to sides of the head in the azimuth plane. This ensures that there is unlikely to be an asymmetry in the amount of masking that may occur at one ear or the other due to head shadowing. Performance under such stimulus conditions most likely depends on binaural processing (localization and/or binaural analysis) for detecting, discriminating, or recognizing targets in the presence of spatially separated maskers.[1]

Binaural processing of interaural time difference (ITD) and interaural level difference (ILD) is the primary mechanism for sound source localization and binaural analysis in the front azimuth plane (e.g., see Blauert, 1997). It has been

[a)]Electronic mail: william.yost@asu.edu

shown that other spatial cues, such as those associated with the head-related transfer function (HRTF), are not normally used for sound source localization in the front azimuth plane (see Shub *et al.*, 2008b). As such, binaural processing, under conditions of a centrally located target with symmetrically placed maskers, is most likely responsible for SRM in the front azimuth plane. The main question addressed in this paper is how does SRM depend on the number of spatially separated maskers when performance can be assumed to be primarily the result of binaural processing and not head shadowing or cues that occur for stimuli that are not on the azimuthal plane?

Yost (1997), Zurek (1993), Bronkhorst (2000), and Litovsky (2012) among others have reviewed the literature on SRM. Yost's (1997) review mainly covered studies of discrimination and recognition as opposed to detection. His review indicated that SRM in recognition and discrimination tasks was smaller than when listeners were asked to detect the presence or absence of a target (signal) in a background of maskers, especially for lateralization/MLD tasks over headphones. He observed that in almost all of the studies he reviewed there were only one or two maskers. Yost (1997) surmised that SRM, for recognition and discrimination, might be larger if there were more maskers representing a more complex auditory scene as compared to one or two maskers because binaural processing might play a larger role.

Zurek (1993) and Bronkhorst (2000) both also reviewed a literature on speech intelligibility in multi-talker conditions that included studies of SRM. Most of the studies in these reviews also involved one or two maskers. When studies involved more than two spatially separated masker sources, some of the masker sources were usually placed asymmetrically about a centrally placed target source. Both reviews included non-SRM conditions. Bronkhorst (2000) reviewed studies involving listeners with hearing impairment. Both Zurek (1993) and Bronkhorst (2000) described models to account for the data they reviewed. Zurek's model was a combination of an Articulation Index (AI, see Zurek, 1993) calculation and a modification of the EC model (see Durlach, 1963), while Bronkhorst's model included separate terms to account for the contributions of multiple symmetrically or asymmetrically placed maskers. These models accounted for a lot of the reviewed data and reflected the trend that SRM decreases as the number of masker sources increases. The models were also able to account for the fact that there is less SRM when masker sources are symmetrically located relative to the target source than when masker sources are asymmetrically located (i.e., when the target-to-masker ratio differs at the two ears due to head shadow).

Litovsky's (2012) review included many of the studies reviewed previously, but added discussion of the role of informational and energetic masking in SRM (discussed below), and contained studies of SRM in special subject populations (e.g., children and cochlear implant patients). Most of the studies reviewed by Litovsky involved one or two maskers and often involved conditions in which masker sources were asymmetrically placed relative to the target source. Thus, a great deal of the literature reviewed by Yost (1997), Zurek (1993), Bronkhorst (2000), and Litovsky (2012) involved masker situations in which the target-to-masker ratio at one ear could be greater than at the other ear due to "head shadow." And these reviews covered aspects of masking (e.g., monaural variables like amplitude modulation) that are not strictly related to binaural processing on its own. However, these reviews presented several studies documenting the role of binaural processing in SRM for (usually) two maskers. What is not clear from these reviews is how SRM that is primarily the result of binaural processing changes when the number of maskers increases beyond two. This, again, is the topic of the current paper.

In addition to these SRM studies, Santala and Pulkki (2011) and Kawashima and Sato (2015) showed that listeners are not able to identify or localize more than about three to five simultaneously presented, spatially separated sources. If SRM depends on processes related to identification and/or localization, then perhaps SRM would be very small, if it existed at all, when the number of maskers exceeded three or four. There are very few data and varying arguments regarding the relationship between SRM and the number of maskers when SRM primarily depends on binaural processing alone. Specifically, there are very few data involving more than two maskers located symmetrically in the front azimuth plane.

In this paper, SRM was measured in the front azimuth plane for three conditions with two, four, or six masker loudspeakers symmetrically located around the center target loudspeaker. The task was speech (single word) recognition. The conditions were condition TsMs in which a speech (s) word uttered by male talkers masked (M) target (T) speech words uttered by a female talker, condition TsMn in which filtered and modulated noise (n) maskers masked target speech words uttered by a female talker; and condition TfsMfs in which filtered speech (fs) words uttered by male talkers masked filtered target speech words uttered by a female talker.

Bronkhorst and Plomp (1992) conducted a study that involved two, four, and six maskers; and in a few conditions these masker sources were placed symmetrically about the centered target source, similar to conditions examined in the present paper. However, the rest of the stimulus conditions were different from those used in the present study. Bronkhorst and Plomp (1992) found that SRM decreased slightly as the number of maskers was increased; the same was found for speech intelligibility in the co-located target and masker conditions. Kidd *et al.* (2016) investigated SRM with two and four maskers when the maskers were spatially separated symmetrically about the target. The main aim of this paper was on the role of energetic and informational masking, but the results did indicate less SRM for four as compared to two maskers. Other studies (e.g., Freyman *et al.*, 2004) have used several maskers (up to ten) in studies of SRM, but when most of the maskers are mixed and presented from one loudspeaker (i.e., presenting a "babble masker" from one loudspeaker and, perhaps, a different "babble masker" from a different loudspeaker). The purpose of the present paper was to study conditions in which

individual masking sounds could each be potentially localized at a different location.

As discussed in the Litovsky (2012) review of the SRM literature, it is often important to consider the type of masking that might be involved, i.e., mainly informational masking, mainly energetic masking, or a combination of informational and energetic masking (see Kidd *et al.*, 2008; Kidd *et al.*, 2016). In the current study, condition TsMs involves speech maskers masking speech targets; a combination of informational and energetic masking since speech-on-speech masking can represent an estimate of informational masking based on the similarity of the target and masker speech stimuli (see Watson, 2005; Kidd *et al.*, 2008; Kidd *et al.*, 2016). Note that the spectral/temporal overlap of the target and maskers in condition TsMs can also produce energetic masking (see Kidd *et al.*, 2008; Kidd *et al.*, 2016) at points where the two speech stimuli temporally and/or spectrally overlap.

Condition TsMn involves speech targets accompanied by noise bursts for maskers that were filtered and modulated to be similar to masker words. Results for condition TsMn are assumed to provide an estimate of primarily energetic masking with little informational masking. This is because the noise maskers present considerable energetic masking due to the temporal/spectral overlap with the speech target while, at the same time, they are not perceived as speech or speech-like and are therefore not perceptually similar to the speech target.

Condition TfsMfs involves differentially filtering the speech targets and maskers as originally suggested by Arbogast *et al.* (2002). The filtering operation reduces the spectral overlap between the speech target and maskers but retains the intelligibility of the words, and as such should reduce the amount of energetic masking, yielding more informational than energetic masking. These assumptions are clearly subject to testing as will be done.

There is a literature that suggests that SRM is larger when there is considerable informational masking as compared to when there is primarily energetic masking (see Freyman *et al.*, 1999; Litovsky, 2012, for a review). Further exploration of this literature will be described in the Discussion section in the context of the data obtained in these experiments. To repeat, the major aim of the experiments of this paper is to determine the extent to which SRM for informational and/or energetic masking depends on the number of maskers when binaural processing is likely the primary process involved.

II. METHODS

A. Listeners

In each of the three conditions there were 18 listeners who participated in only one condition each (54 total listeners). In condition TsMs there were 12 females and 6 males all between 19 and 34 years of age. In condition TsMn there were 15 females and 3 males all between 19 and 34 years of age. In condition TfsMfs there were ten females and eight males all between 19 and 34 years of age. All listeners in all conditions were American English speakers and reported normal hearing. All procedures were approved by the Arizona State University Institutional Review Board for the Protection of Human Subjects.

B. Stimuli

In all conditions, sounds were generated via a 24-channel Digital-to-Analog (DA) converter (two, Echo Gina 12 DAs, Santa Barbara, CA) at a sample rate of 44 100 samples/s/channel.

In all conditions the target speech word was randomly chosen from twelve, one-word country names (Belgium, Britain, Burma, China, Cuba, Japan, Korea, Libya, Mexico, Norway, Russia, Turkey) spoken by one of six randomly chosen female speakers. The 72 target words (12 country names by six female talkers) were not all originally equally recognizable in a pilot experiment (five listeners) with co-located target and masker with the broadband noise presented at 65 dBA and each target word presented at 62 dBA (−3 dB target-to-noise ratio). Most words were recognized at about a mean of 0.80 accuracy. Words that were recognized at accuracies greater than 0.90 or lower than 0.65 were regenerated and retested. After several iterations of sampling words, all 72 target words were recognizable at between a mean of 0.65 and 0.90 accuracy at the same target-to-noise ratio and overall stimulus level noted above. Within this range there was variability in word recognition among the five pilot listeners, suggesting that some of the differences in the ability to recognize target words is due to the listeners, and not just to the word being uttered.

In condition TsMs, maskers were words randomly chosen from the same 12 one-word country names used for the targets. Maskers were spoken by one of six randomly chosen male speakers. The same pilot-study procedure described above for target words was used to finalize the masker words so that each of the 72 masker words was recognized in broadband noise between 0.65 and 0.90 proportion correct masker words. All target and masker speakers were native users of English. All words were temporally centered in the middle of 750-ms duration files, and were filtered from 125 to 8000 Hz with a three-pole Butterworth filter implemented in MATLAB. All words were generated at the same root-mean-square (rms) level. Whenever multiple maskers were presented (either co-located with the target at the center loudspeaker or spatially separated from the target and from each other), the rms level of each masker word was always the same and the overall level of the combined maskers was kept at 65 dBA, measured at the position of the listener's head. For each masker combination, three target levels were obtained so that there was a low target-to-masker ratio in decibels, determined in pilot work, that yielded proportion of words correctly recalled near 0.25 (chance is 0.0833 = 1/12 possible country names). This target level was the lowest of three target-to-masker ratios (dB), with the other two more intense target levels yielding target-to-masker levels that were 4 and 8 dB greater, so that the target-to-masker ratios differed by four decibels. Thus, the summed overall level of the maskers, no matter the combination, was always the same (65 dBA), and changes in target level provided

three target-to-masker ratios for estimating three-point psychometric functions. The lowest target-to-masker ratio was intended to yield approximately 0.25 proportion correct target word recognition performance, and the highest target level would not yield proportion correct performance near 1.0 (i.e., the three target-to-masker ratios would be near the midpoint of the psychometric function where the psychometric function is linear or nearly so).

In condition TsMn the maskers were filtered and amplitude modulated 750-ms noise bursts (20-ms cosine-squared rise-fall times). A complete Fast Fourier Transform (FFT) of each of the 72 masker words was calculated and then the component amplitudes were normalized with the most intense component for each word having an amplitude of 1.0. On each trial, 750-ms independent noise bursts were generated in the frequency domain such that the amplitudes of each spectral component between 125 and 8000 Hz (same bandwidth as the words) were approximately Rayleigh distributed (using the method suggested by Hartmann and Pumplin, 1988). The starting phases were randomly and uniformly distributed from 0 to 2pi. The amplitude of each noise spectral component was multiplied by the amplitude values obtained from the FFT of each word. Then the resulting amplitude and phase spectra were inversed Fourier transformed back to the time domain yielding nearly Gaussian distributed instantaneous amplitudes (see Hartmann and Pumplin, 1988) spectrally shaped (based on word spectra) noise bursts. The Hilbert envelope of each word was then extracted on each trial and multiplied by the appropriate filtered noise burst, yielding a time-domain waveform with the overall amplitude modulated envelope of the word. The same procedure described for condition TsMs was used for condition TsMn to generate equal masker levels and the three different target-to-masker ratios (dB) for each masker combination. Informal listening by five listeners indicated that the filtered/modulated noises were not perceived as speech or even speech-like. All five listeners indicated that many noises were perceived as having some subtle temporal fluctuation and some were judged to be "brighter" than others. Although these noise bursts were not perceived as speech, they had the long-term spectral contour and amplitude-modulated envelope of the words.[2]

In condition TfsMfs, a filtering process like that suggested by Arbogast *et al.* (2002) was used. In this condition the same female target words and male masker words as were used in condition TsMs were filtered using a two-equivalent rectangular bandwidth (ERB) wide Gammatone filter bank (center frequencies from approximately 125 to 8000 Hz).[3] The target words were filtered with every odd filter in the filter bank and the masker words with the even filters with each masker word filtered in the same way. Figure 1 indicates the outputs of the filter bank for a broadband noise, with the filtered noise shown in black being the consequence of the filtering used for the target words and the lighter shade of gray the consequence of the filtering used for the masker words. Informal listening suggested that the filtered words were about as intelligible as the original unfiltered words, although no formal tests of speech intelligibility were conducted. The same procedures as were used in conditions TsMs and TsMn were used to generate equal masker levels and three appropriate target-to-masker ratios (dB) in condition TfsMfs.

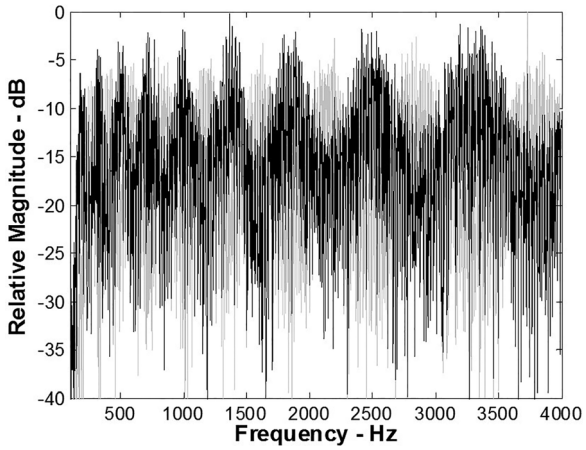

FIG. 1. The output of the two-ERB wide filter bank used to filter the target and speech words in condition TfsMfs is shown. A broadband noise was used. The dark areas represent the output of every odd filter which was used to filter target words, while the lighter areas are the outcomes for the even filters used to filter masker words.

C. Listening environment

The same listening room used in Yost *et al.* (2015) was used in the present experiments. The room is a 10 ft × 15 ft × 10 ft lined on all six surfaces with acoustic foam. The wideband reverberation time (RT60) is 102 ms and the ambient noise level is 32 dBA. Twenty-four loudspeakers [Boston Acoustics 100 × (Peabody, MA)] are on a 5-ft radius circle (i.e., azimuth array with 15° loudspeaker spacing) at the height of listener's pinna. There is a control room from which listeners are monitored by an intercom and camera. Listeners were instructed to face the center loudspeaker, which had a red dot on it, at all times. Listeners were monitored on each trial and rarely failed to face the center loudspeaker. In these rare cases, the trial was repeated.

D. Procedure

The target word, spoken by a female, was always presented from the center loudspeaker in all three conditions. The masker words, spoken by males, were presented from the following source locations in the three conditions (see Fig. 2).

Condition TsMs (see Fig. 2): For two masking sources (2-m), there were four possible masker source locations with masker sources placed symmetrically about the center loudspeaker at ±30°, ±45°, ±60°, or ±90°. For four masking sources (4-m), two masker source combinations were used with the masker sources placed symmetrically about the center loudspeakers at ±(30° and 60°); or ±(45° and 90°). For six masking sources (6-m), masker sources were placed at ±(30° and 60° and 90°) relative to the center loudspeaker.

Conditions TsMn and TfsMfs (see Fig. 2): For two masking sources, only the combination in which the masker sources were placed at ±90° relative to the center

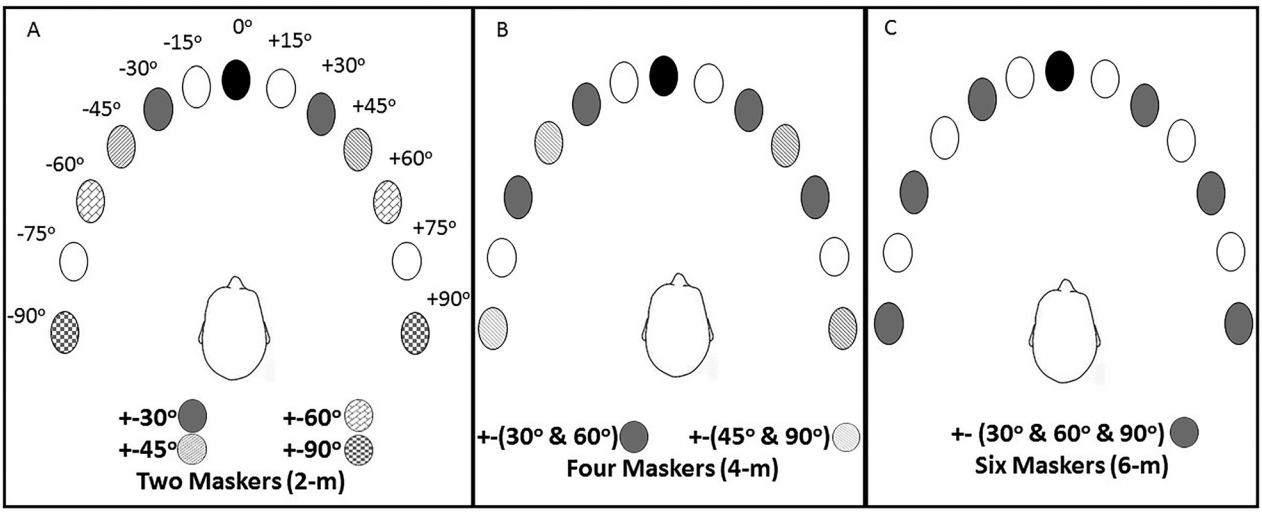

FIG. 2. The location of the loudspeakers used for the target (center dark filled circle) and the maskers (off center circles-different filled circles for different loudspeaker combinations) for the (A) two (2-m on the left), (B) four (4 -m in the middle), and (C) six (6-m on the right) masker cases. No sound was presented from loudspeakers located at ±15° or ±75° in any case (unfilled circles).

loudspeaker was tested. For four masking sources, only the combination in which the masker sources were placed at ±(45° and 90°) relative to the center loudspeaker was tested. For six masking sources, masker sources were placed at ±(30° and 60° and90°) relative to the center loudspeaker (as they were for condition TsMs).

For each of the three number of masker sources (two, four, and six maskers) and for each of the three conditions (TsMs, TsMn, and TfsMfs), a co-located condition was tested in which the female target and male masker words were mixed and all presented from the center loudspeaker. The same procedure as described for the spatially separated target-masker source conditions was used to determine equal masker rms levels and target-to-masker ratios for these co-located conditions.

The 12 country names were displayed on the response terminal on each trial and listeners indicated which word they perceived as being uttered by the female target presented from the center loudspeaker. No feedback was provided. On each trial a randomly determined word from a randomly determined female talker was presented. Masker words were similarly randomly determined, but all masker words were different on each trial and no masker word was the same as a target word on a trial (i.e., all words were different on each trial). Twenty-four trials were used to obtain each point on each of the three-point psychometric functions (72 trials/psychometric function × 22 psychometric functions/listener × 18 listeners yields 28 512 total trials per each of the three conditions or 85 536 total trials across the three conditions). Each of the 72 target words (12 words by 6 female talkers) occurred once per psychometric function in random order. Thus, any differences between psychometric functions for any listener could not be attributed to a different set of target words being sampled for one psychometric function as compared to another. Since there was always more than one masker per trial, more than 72 masker words were used for each psychometric function. The only additional restriction (see above) placed on randomly sampling masker words was that all 72 possible masker words had to be used at least once per psychometric function.

For each of the three conditions (TsMs, TsMn, and TfsMfs), listeners started with one set of the 72 target words in the co-located case at the highest target-to-masker ratio as a practice session. Then the co-located and spatially separated cases were presented in random order (starting with a case that was not used for practice) for each listener over the course of that condition. A complete psychometric function for one target and multiple masker case was obtained (in three, 24-trial blocks) and then the next case to test was determined randomly so each case occurred once.

III. RESULTS

Figure 3 shows the main results as the mean (18 listeners per condition) proportion correct of reported words and plus/minus one standard deviation as a function of target-to-masker ratio (dB). The columns in Fig. 3 represent the three combinations of target and masker sources (two, four, and six maskers), and the rows represent the three conditions (TsMs, TsMn, and TfsMfs). The dark solid line and solid squares in each panel represent the co-located target and masker cases. The lighter colored lines and non-squared symbols represent the spatially separated target/masker sources as indicated in the legend in each panel.

The attempt in a pilot experiment to find target-to-masker ratios that were the lowest ratio for each situation that might yield a proportion correct of about 0.25 was fairly successful. In all cases the entire psychometric function was within the bounds of 0.15 to perfect performance (1.0), which was the main aim of the pilot experiment (in 89% of the cases proportion correct responses were between 0.23 and 0.85).

It appears as if all of the psychometric functions are nearly linear on the coordinates of proportion correct target words versus target-to-masker ratio in dB. The slopes of the psychometric functions also appear to be about the same.

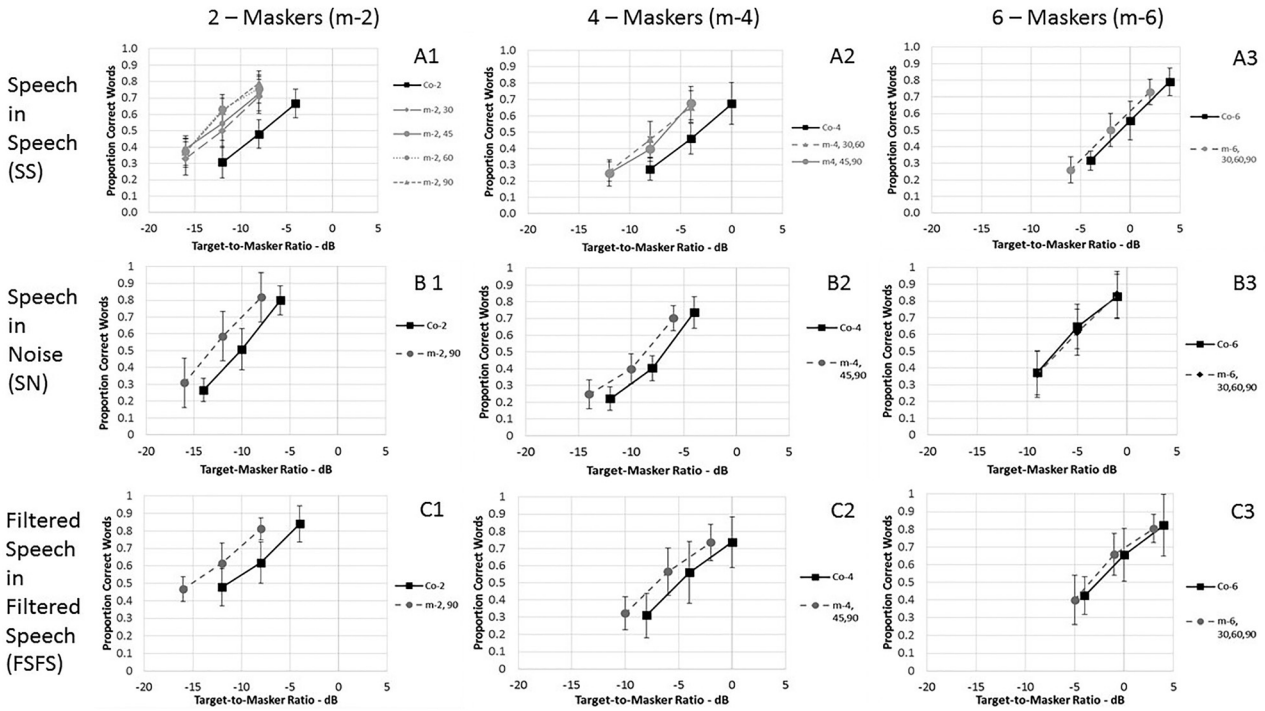

FIG. 3. Psychometric functions for each of nine cases are shown. Three rows (A, B, C) represent conditions TsMs, TsMn, and TfsMfs from top to bottom. Three columns (1, 2, 3) represent number of maskers (two, four, and six from left to right). Data are based on mean proportion of correct responses (18 listeners per condition) as a function of target-to-masker ratio in dB. Vertical error bars are ± one standard deviation. Dark squares and dark solid lines represent the co-located cases, while the lighter color, non-squared data represent the spatially separated cases as indicated in the legend of each figure (see Fig. 2).

The average slope across all psychometric functions (when fit with straight lines) was 0.048 proportion words correct/dB of target-to-masker ratio with a standard deviation of 0.008 proportion words correct/dB of target-to-masker ratio. A one-way mixed analysis of variance (ANOVA) was calculated for the slopes of all of the psychometric functions. There was no statistically significant difference in slope at a 0.05 level of significance.

For Fig. 4, the data of Fig. 3 were normalized assuming that all of the data could be fit with straight-line psychometric functions. That is, target-to-masker ratios (dB) for each data point were multiplied by the slope of the line that best fit those data to obtain a normalized proportion correct word estimate. The best fitting line to these normalized data (in Fig. 4) was obtained using linear regression (slope = 0. 0476 proportion correct/dB) and this best fitting line accounted for over 97% of the variance of the linear fit. Thus, for the rest of the data analysis, it is assumed that all of the psychometric functions are linear. The average slope of these psychometric functions was 0.048 proportion correct/dB, and further data analysis will be based on the best fitting line to each psychometric function.

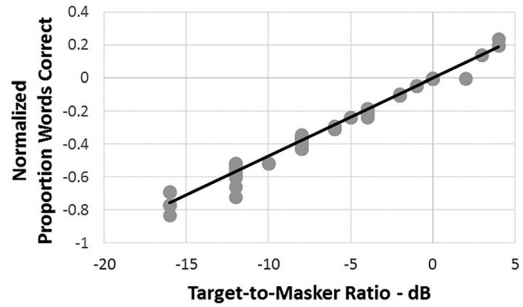

FIG. 4. Normalized proportion correct responses (see text for explanation) as a function of target-to-masker ratio in dB for the 22, three-point psychometric functions (66 data points are plotted) shown in Fig. 3. Best fitting regression line is shown. It has a slope of 0.0476 proportion correct/dB which accounts for more than 97% of the variance of the linear fit to the data.

In Fig. 3 SRM is associated with the separation of the psychometric functions when masker sources are spatially separated from the target source versus when the target is co-located with the maskers. Figure 5 displays SRM in terms of target-to-masker ratio (dB) required for a proportion of correct target words reported of 0.5417 (half way between chance performance of 0.0833 proportion correct and perfect performance of 1.0). The determination of the target-to-masker ratio (in dB) was made on the basis of the best fitting linear psychometric function for each condition. SRM, shown in Fig. 6 is measured as the difference in proportion of correct words for spatially separated target-masker source cases versus co-located cases. For Fig. 6, the proportion of correct words was calculated from the intercepts of the best fitting linear psychometric functions (since all psychometric functions had statistically the same slope) for each condition (see above). In general, SRM, measured either in terms of target-to-masker ratio (Fig. 5) or proportion correct words (Fig. 6), decreases as the number of masker sources increases from two, to four, to six. SRM is largest for condition TsMs, next largest for condition TfsMfs, and usually smallest for condition TsMn. In condition TsMs, the amount

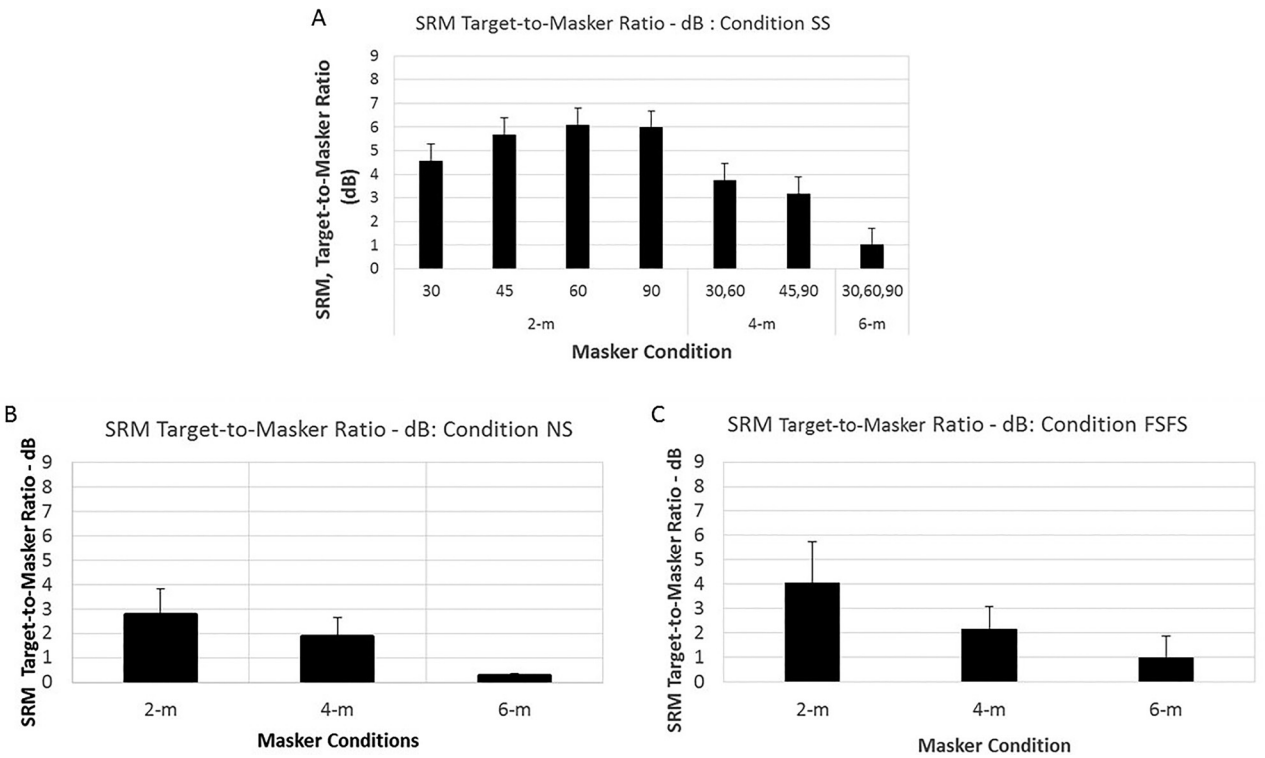

FIG. 5. Mean (18 listeners for each condition) SRM expressed in target-to-masker ratio (dB) required for a proportion correct of 0.05417 is shown as a function number of maskers (2-m, 4-m, 6-m) and spatial separation of the masker sources from the target source for each number of maskers (shown in the legends). Data for (A) condition TsMs are shown on top, (B) condition TsMn on the lower left, and (C) condition TfsMfs on lower right. Error bars are ± one standard deviation.

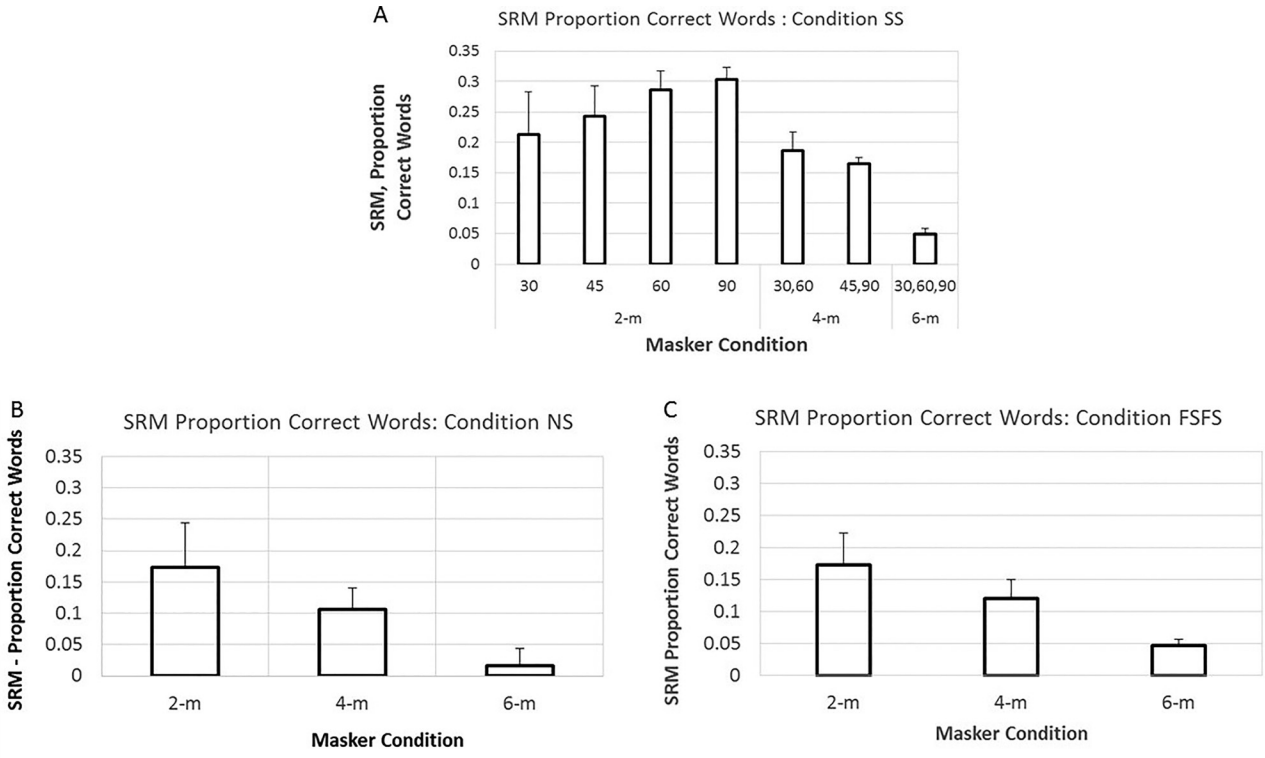

FIG. 6. Mean (18 listeners for each condition) SRM expressed as proportion correct responses derived from the slope of the psychometric functions (see text) is shown as a function number of maskers (2-m, 4-m, 6-m) and spatial separation of the masker sources from the target source for each number of maskers (shown in the legends). Data for (A) condition TsMs are shown on top, (B) condition TsMn on the lower left, and (C) condition TfsMfs on lower right. Error bars are ± one standard deviation.

of spatial separation of the masker loudspeakers from the target loudspeaker appears to have a small effect on SRM as measured in the two, but not for the four masker cases.

Figure 7 shows individual data for one listener for the two-masker configuration in condition TsMs in the co-located (dark symbols) and the 90° spatially separated configuration (gray symbols). The lines are the best fitting lines (over 97% of the variance accounted for in each case) to the data (slope of 0.046/dB for the co-located data and 0.049/dB for the 90° separation data). These individual data are illustrative of the data obtained for all listeners across all experimental cases.

The data of Fig. 5 (SRM in terms of target-to-masker ratio in dB) were analyzed with ANOVA statistics and subsequent *post hoc* t-tests. For each condition (TsMs, TsMn, and TfsMfs) a repeated measures ANOVA was performed with number of maskers and spatial separation as the two main variables for condition TsMs and just the number of maskers for conditions TsMn and TfsMfs. In all tests the significance criterion was 0.05. For condition TsMs both main effects (number of maskers and spatial separation) were statistically significant. There was also a statistically significant interaction between number of maskers and spatial separation. *Post hoc* t-tests indicated that only the 30° separation condition differed significantly from the other spatial separation cases when there were two maskers. The two different spatial separation conditions for the four masker cases in condition TsMs were not statistically significantly different. For conditions TsMn and TfsMfs there was a statistically significant difference due to the number of maskers. A pair-wise, *post hoc* t-tests for condition TsMn indicated that there were statistically significant differences among all three number of masker cases. For condition TfsMfs all pair-wise, *post hoc* t-tests yielded statistically significant differences based on number of maskers.

To obtain a statistical estimate of the difference due to condition, mean SRM as measured for target-to-masker ratio (dB) was computed for each of the three conditions (TsMs, TsMn, TfsMfs) and those three means were analyzed by a between groups ANOVA. There was a statistical main effect (0.05 level of significance) based on the three conditions, and subsequent pair-wise, *post hoc* t-tests indicated that there were statistically significant differences among all condition comparisons.

Table I indicates the proportion of target word errors that were masker words (i.e., the proportion of errors when the target words were mistaken as masker words) for two conditions (TsMs and TfsMfs) and the three combination of number of maskers. Proportions were calculated across all masker/target cases including different masker source locations and the co-located cases. Collection of these data was only appropriate for conditions TsMs and TfsMfs in which the maskers were words. The proportion of the target words errors that were mistaken as masker words decreased as the number of maskers increased and were slightly lower for condition TfsMfs than for condition TsMs.

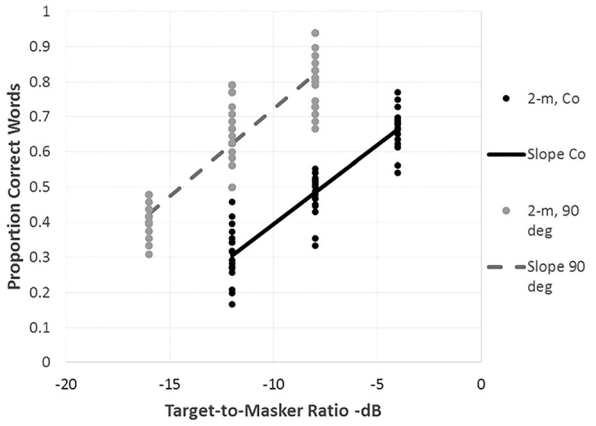

FIG. 7. One listener's data (which are similar to all other listeners' data) for condition TsMs in the co-located case (2-m, Co: dark symbols) and the 90° spatial separation case (2-m, 90°: gray symbols), in which proportion of correct words is plotted as a function of target-to-masker ratio in dB (54 data points per function). The best fitting regression lines are plotted for each case (Slope Co = 0.046 proportion correct/dB, and Slope 90° = 0.049 proportion correct/dB).

IV. DISCUSSION

The major finding of this study is that SRM, independent of the type of masking (energetic and/or informational as determined by the three different conditions), decreases as the number of spatially separated masker sources increases from two to six. It appears as if there is almost no SRM for six masker sources. Put another way, when SRM is primarily the result of binaural processing alone it appears to be limited to a simple auditory scene with five or fewer sound sources. The amount of SRM obtained in this study is within the range of that found in other studies in which symmetrically placed masker sources were used, although there are very few studies involving more than four symmetrically placed masker sources (e.g., for two maskers, the comparisons are to Bronkhorst and Plomp, 1992; Srinivasan *et al.*, 2016; Kidd *et al.*, 2016; and four and six maskers comparisons to Bronkhorst and Plomp, 1992; Kidd *et al.*, 2016). The findings in the present paper were obtained for word recognition when maskers were male speakers at sources located in the front azimuth plane and placed symmetrically around a centered female target speaker. It is not certain if similar results would be obtained for other measures of performance (e.g., detection), other stimuli (e.g., sentences or if the masker and target sounds were from talkers of the same gender), or other masker source arrangements (e.g., maskers located asymmetrically about the target, behind the listener, or at different elevations or distances). It should also be recognized that the experiment took place in a reflection-reduced room and, as such, the role of reflections/reverberation was reduced. The literature suggests that speech intelligibility is dependent on reverberation (e.g., see Bronkhorst, 2000) as is SRM (e.g.,

TABLE I. Proportion of reported target word errors that were masker words as a function of the number of maskers (2-m, 4-m, and 6-m) and condition (TsMs and TfsMfs).

	2-m	4-m	6-m
TsMs	0.82	0.54	0.18
TfsMfs	0.77	0.49	0.12

see Kidd *et al.*, 2008), so it is uncertain as to what the outcomes of these experiments might have been if testing had been done in a space with a different amount of reverberation.

The assumption put forth in Sec. I was that, with symmetrically placed maskers in the front azimuth field, SRM would be based on binaural processing. This binaural processing could be due to either sound source localization of the target and masker sources and/or binaural analysis of auditory spatial cues associated with the spatially distributed target and maskers sources. It is hard to image that some other auditory process, besides binaural processing, played a major role in the SRM measured in this paper. The answer to the question as to what type of binaural processing (sound source localization and/or binaural analysis) was used is not as clear. In answering this question one would probably have to consider the type of masking that was involved (energetic and/or informational masking).

In Sec. I, a distinction was made between informational masking (condition TfsMfs), energetic masking (condition TsMn), and combined informational and energetic masking (condition TsMs). Several analyses were performed to investigate these assumptions about informational and energetic masking for the conditions of this paper. In Fig. 8 on the top row, the co-located (Co) data are compared as a function of the number of maskers (two maskers: Co-2, four maskers: Co-4, and six maskers: Co-6) for each of the three conditions (condition TsMs left, condition TsMn middle, and condition TfsMfs right of Fig. 8). Recall that the rms level of the total masking energy arriving at the position of the listener's head is the same for all of the data in Fig. 8. Thus, any changes in performance with an increased number of maskers is not due to the level of the overall masker stimulus changing.

For Fig. 8 in the top row considering condition TsMs there is a clear increase in target level required for word recognition as the number of maskers increases. This is most likely due to the difficulty in attending to a target word in the background of an ever increasing number of competing words (masker words). This explanation would be consistent with the concepts of informational masking (see Brungart, 2001; Ihlefeld and Shinn-Cunningham, 2008; Kidd *et al.*, 2008; Kidd *et al.*, 2016). Bronkhorst's (2000) review and Bronkhorst and Plomp (1992) report a similar decrease in speech intelligibility as the number of maskers in co-located cases increases.

For condition TsMn in the co-located case (Fig. 8, top row), if energetic masking were determined entirely by the rms level of the masker, then there should be no differences in the psychometric functions as the number of maskers increases. This appears to be true in comparing the four versus six masker case, but not in comparing the four or six masker to the two masker case. The difference in performance between the two masker cases and the four (or six) masker cases might be due to stimulus (energetic masking) variables other than overall rms masker level. These stimulus variables might allow for a form of "better-ear listening" if a word from one masker did not mask the target word as much as the word from a different masker. Because the target word would only be affective by one masker, that masker would be asymmetrically located allowing for a form of "better ear listening." That is, the spectral and/or temporal (e.g., envelope, and the use of "glimpsing," e.g., see Brungart and Iyer, 2012) differences (that existed in condition TsMn) when there are two as compared to four noise maskers might make it easier to recognize target words in the two masker cases when the spectral/temporal differences between the target and masker words are likely to be larger than for the four masker cases. That is, as the number of maskers that are added together increases, the spectral and modulation patterns due to any one masker will become

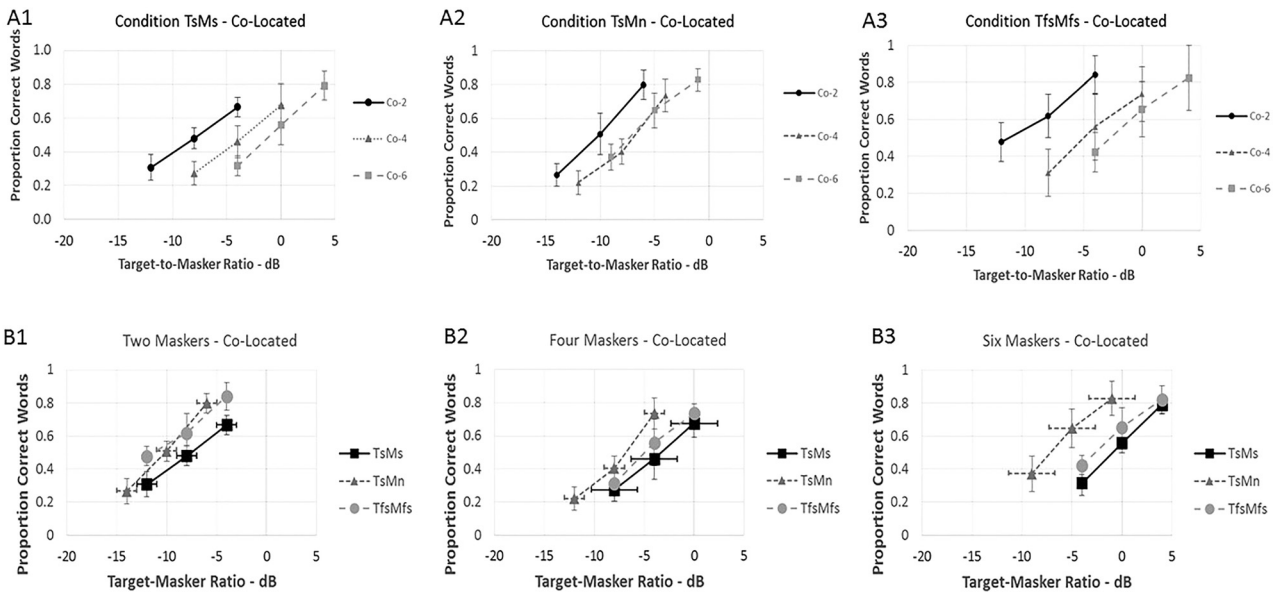

FIG. 8. Top Row (A1-A3): Psychometric functions for the co-located cases for each of the three conditions [condition TsMs left (A1), condition TsMn middle (A2), and condition TfsMfs right (A3)], and for each number of maskers (Co-2, Co-4, and Co-6). Error bars are ± one standard deviation (see Fig. 3). Bottom Row (B1-B3): Psychometric functions for the co-located cases for each number of maskers (Co-2 left (B1), Co-4 middle (B2), Co-6 right (B3)), and for each of the three conditions (condition TsMs, condition TsMn, and condition TfsMfs). Error bars are ± one standard deviation (see Fig. 3).

reduced due to the addition of the other maskers. Thus, the data in condition TsMn seem somewhat consistent with an argument that masking is largely energetic in condition TsMn. However, there may be a very small effect due to spectral/temporal differences between the maskers and the target when comparing two versus four maskers, but probably not in comparing four versus six maskers.

There is a separation among the co-located psychometric functions in Fig. 8 for condition TfsMfs similar to that measured for condition TsMs, and there is less SRM for condition TfsMfs than for condition TsMs. These results appear to be consistent with the argument posed above for condition TsMs regarding decreased speech intelligibility as the number of maskers increases. So condition TfsMfs might indeed represent mostly informational masking, since there was reduced energetic overlap of the maskers and targets due to the differential filtering in condition TfsMfs.

The bottom row of Fig. 8 reinforces the interactions shown in the top row of Fig. 8. In the bottom row of Fig. 8 the data are replotted for the co-located cases for each number of maskers (two maskers left, four maskers middle, and six maskers right) as the proportion of correct words as a function of target-to-masker ratio (dB) for each of the three conditions (TsMs, TsMn, and TfsMfs). In all cases, condition TsMs indicates poorest performance, with condition TfsMfs indicating similar, but slightly better, performance than condition TsMs, and condition TsMn indicates best performance.

Another type of comparison regarding energetic and informational masking is provided in Fig. 9. In Fig. 9 the amount of co-located masking was determined in terms of the target-to-masker ratio (dB) required for 0.5417 proportion words correct obtained from the best fitting linear psychometric functions ("Threshold Masking"). Figure 9 shows "Threshold Masking" for the sum of conditions TsMn plus TfsMfs (back bar) as compared to that obtained in condition TsMs (white bar) for two, four, and six maskers. The fact that summed "Threshold Masking" for conditions TsMn plus TfsMfs is close to that obtained for condition TsMs is consistent with the assumption that masking in condition TsMs is energetic masking (condition TsMn) plus informational masking (condition TfsMfs, see Arbogast *et al.*, 2002, for other issues involved with a similar analysis).

The data of Table I also seem consistent with an assumption that as the number of masker words increases it is more and more difficult to identify a masker word in the mixture of several different words presented simultaneously (as reported by Kawashima and Sato, 2015). As a consequence, a missed target word is less and less likely to be reported as a masker word as the number of maskers increases. It is not clear as to why the proportions reported in Table I are lower for condition TfsMfs than for condition TsMs, except to note that there were different listeners for the two conditions, and there might have been some slight lowering of target word intelligibility in condition TfsMfs due to the filtering as compared to condition TsMs. Others have reported that it is often the case that a high proportion of missed target words are those of the masker (e.g., see Brungart, 2001; Kidd *et al.*, 2008). Only one or two maskers have been used in these studies as compared to the four and six masker cases of the present study.

In regard to SRM, an analysis like that performed for Fig. 9 is displayed in Fig. 10 in order to deal with issues of how SRM changes due to the type of masking (energetic and/or informational). Figure 10 shows SRM in terms of target-to-masker ratio (dB) as a function of the number of maskers (see Fig. 5). The amount of SRM in condition TsMn (energetic masking, black section) plus that for condition TfsMfs (informational masking, gray section) is compared to SRM for condition TsMs (energetic and informational masking, white bar). Only the $\pm 90°$ separation cases for two maskers and the $\pm(45°$ and $90°)$ cases for four maskers were used for condition TsMs in Fig. 10. SRM was computed for condition TsMn and TfsMfs. Figure 10 indicates that SRM for condition TsMs in which presumably both energetic and informational masking exist is about equal to the sum of SRM for condition TsMn (mainly energetic masking) plus condition TfsMfs (mainly informational masking). We have no theoretical point to make about these relationships, but point out that the amount of SRM summed as indicated in Fig. 10. While the analyses in Figs. 8–10 do not unequivocally establish the assumptions that were made about informational and energetic masking and their combination, they are largely consistent with these assumptions.

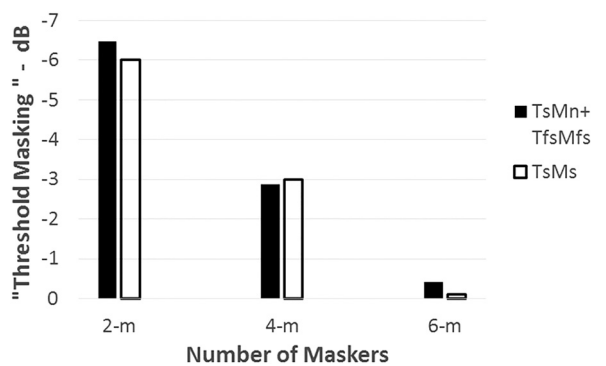

FIG. 9. "Threshold Masking" in dB (see text) as a function of the number of maskers (2-m, 4-m, and 6-m) for the sum of conditions TsMn and TfsMfs (black bar) compared to condition TsMs (white bar). Data are for the co-located cases.

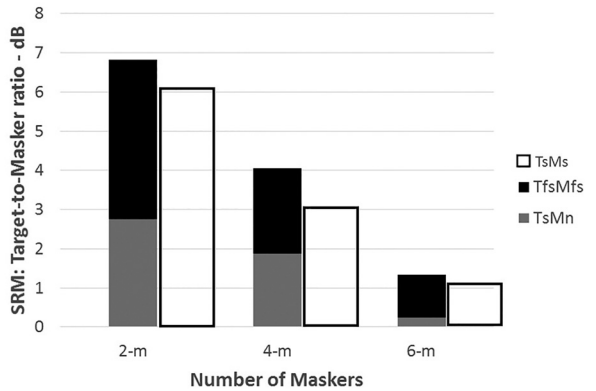

FIG. 10. SRM in dB as a function of the number of maskers (2-m, 4-m, and 6-m) for the sum of conditions TsMn (gray bar) and TfsMfs (black bar) compared to condition TsMs (white bar).

As described in Sec. I, Bronkhorst and Plomp (1992) investigated SRM when there were as many as six maskers and in one condition the six maskers were symmetrically spaced. Their data suggest that the amount of SRM decreased as the number of maskers increased to six, but they measured more SRM for their six-masker case than was measured in the current paper. There are many differences between the Bronkhorst and Plomp (1992) study and the current one. For instance, in the Bronkhorst and Plomp (1992) study, two of the six maskers were behind the listener, target sentences were used, the spatial conditions were simulated over headphones using artificial head recordings, and only noise masker conditions were used (the noises were filtered and modulated based on speech but in a different way than the noises used in the present study). It is not clear which variables may be important to explain differences between the results of the present study and those of Bronkhorst and Plomp (1992).

The data from the experiments of this paper suggest that the slope of the psychometric function relating proportion of correct words to target-to-masker ratio in dB is about the same for all three conditions. That is, the slope of the psychometric function does not vary depending on the type of masking (e.g., energetic, informational, or both). While the average slope of 0.048 proportion correct/dB is within the range of those obtained in other studies (see review in Arbogast *et al.*, 2002), many other studies demonstrated a difference in psychometric function slopes among energetic, informational, and both forms of masking (again reviewed in Arbogast *et al.*, 2002). It is not clear why the data of the present study suggest no difference in slope as a function of the type of masking. Possible reasons for the discrepancy in the data across studies might be the differences in stimuli (e.g., words in the present study versus sentences in most other studies) or the fact that in the present study multiple, symmetrically placed maskers were used as opposed to usually single maskers in the other studies. It might also be the case that estimating the slope of the psychometric function with only three points (especially points near the middle of the psychometric function), as was done in the present study, does not allow for a refined enough measure of slope to determine slope differences. In any case, a possible short coming of this paper in regard to dealing with issues of informational and energetic masking is the limited number of points (three) used to estimate psychometric functions. The main goal of this paper was to better understand how the number of symmetrically spaced maskers affects SRM. Given that the psychometric functions appear to do a reasonable job of approximating performance near the midpoint of the psychometric functions, the measures of performance as a function of the number of maskers appear to be valid.

There is a literature that argues that speech-on-speech masking such as measured in the TsMs condition can be largely due to informational masking rather than energetic masking (e.g., Brungart, 2001; Brungart *et al.*, 2001; Ihelfeld and Shinn-Cunningham, 2008; Kidd *et al.*, 2016). The data support this claim when the maskers and target sounds are spoken by the same person or persons of the same gender. The evidence exists for sentences and when the maskers and targets or collocated. The implication that there is a great deal of informational masking in speech-on-speech masking is often based on the psychometric functions obtained in these studies. As a consequence it might be that in some cases listeners use something like "glimpsing" (e.g., see Brungart, 2001) to pay attention to the softer target in order to determine the talker words. The present study used single words when the masker and talker words were uttered by different gender talkers, and measurements were made when the maskers were spatially separated from the target. Using different gender talkers and spatial separation are very likely to reduce the amount of informational masking compared to co-located target and masker conditions when the talkers and masker words were spoken by the same person or persons of the same gender. As previously mentioned, using 750-ms words rather than sentences should significantly reduce the role of "glimpsing" in processing target words in the presence of masker words. Finally, using only three-point psychometric functions reduces the chance of finding the differences in the psychometric functions reported in the literature mentioned above. Thus, it is probably the case that the role of informational masking is less in the TsMs conditions of this study than in the other studies. It is also probable (as argued above) that "glimpsing" plays only a small role in only in the two-masker conditions of this study. That is, many of the issues discussed in the literature (e.g., Brungart, 2001; Brungart *et al.*, 2001; Ihelfeld and Shinn-Cunningham, 2008; Kidd *et al.*, 2016) may play only a small role in the present study, if they play any role at all.

The data displayed in Fig. 10 are consistent with the findings from other studies indicating that there is more SRM when there is both informational and energetic masking than when there is just energetic masking (e.g., see Brungart *et al.*, 2001). The data of Fig. 10 (as well as those in Figs. 5 and 6) also suggest that SRM is somewhat larger for conditions that are primarily informational masking (i.e., condition TfsMfs) compared to when there is primarily energetic masking (i.e., condition TsMn). Arbogast *et al.* (2002) found a similar effect, but in their study only one masker was used, which means that both binaural processing and the effects of head shadow probably influenced their results, whereas in the present study probably only binaural processing was probably used. The difference in SRM between informational and energetic masking conditions was larger in the Arbogast *et al.* (2002) study as compared to that measured in the present experiment, consistent with the finding that asymmetrical masker placement yields larger amounts of SRM than symmetrical placement, most likely due to head shadow effects (e.g., see Bronkhorst and Plomp, 1992; and the review by Bronkhorst, 2000).

There is a literature that suggests that SRM, when there is informational masking (e.g., conditions TsMs and TfsMfs), is based (entirely or partially) on selective attention (Freyman *et al.*, 1999; Shinn-Cunningham, 2008) in which listeners attend to the target and/or disregard the maskers in order to deal with the similarity of the speech sounds from the two sources. The ability of the listener to locate the target source at a different location than the masker source facilitates such a selective attention strategy (e.g., see Arbogast

et al., 2002; Kidd *et al.*, 2008; Kidd *et al.*, 2016). Binaural analysis processes (e.g., an EC process) that enhance target representations relative to masker representations (e.g., Freyman *et al.*, 1999) may reduce energetic interference among masker and target sounds in SRM conditions (i.e., as in conditions TsMs and TsMn). Such binaural analysis occurs because spatial separation of target and masker sources differentially changes the interaural differences associated with each source. That is, binaural analysis contributes to SRM when there is energetic masking, and sound source localization contributes to selective attention allowing for SRM when there is informational masking.

To the degree that sound source localization plays a role in SRM, it is the ability to locate the various target and masker sources that are producing sound all at the same time that is important (i.e., in SRM studies at least two sources, target and masker, produce simultaneous sound). Very little is known about sound source localization of multiple sound sources producing simultaneous sound (see Yost and Brown, 2013, for a review). Listeners can locate two sources producing simultaneous sound but not as well as one sound source (Yost and Brown, 2013). Recently, Santala and Pulkki (2011) and Kawashima and Sato (2015) studied sound source identification and localization when multiple speech, noise, and/or tonal sound sources (two or more) were used. These studies showed that listeners' ability to identify and localize spatially separated speech words decreased with increasing number of sources and was at chance performance when there were more than four or five spatially separated sources producing words. Thus, it appears as if part of the reason for a decrease in SRM with increasing number of masking sound sources is due to an inability to accurately localize the sources of more than five or so simultaneously presented speech sounds. These result suggest that the auditory scene is small. While a model of attention that can predict SRM in conditions of informational masking has not been developed (but see Lutfi *et al.*, 2013), the data of this paper suggest that such a model would need to link sound source localization accuracy to the prediction of SRM, especially when the number of sound sources is two or more.

Several models of SRM have been proposed and tested. Since the present study was done in a sound field, only models of such conditions will be discussed. That is, a review of models of lateralization and the MLD when stimuli are presented over headphones will not be discussed (see Durlach and Colburn, 1978). Zurek (1993), Bronkhorst (2000), Jones and Litovsky (2011), and Cosentino *et al.* (2014) have all proposed SRM models. The models are of two types: descriptive (e.g., Bronkhorst, 2000; Jones and Litovsky, 2011) and process-based (Zurek, 1993; Cosentino *et al.*, 2014). The process-based models have used many aspects of Durlach's (1963) Equalization-Cancellation (EC) model and other proposed neural processing schemes (e.g., Costention *et al.*, 2014). Zurek (1993) also included aspects of Articulation Index calculations along with the EC model. Applying the EC model or the other aspects of the process-based models requires calculations based on the actual stimuli presented to listeners. Recording the actual stimuli at the ears of the listeners was not done in the present experiment, so these process-based models cannot be thoroughly tested for the conditions of this paper.

The Bronkhorst (2000) model contains two main descriptive terms: one for conditions when maskers are asymmetrically placed (i.e., effects of head shadow can occur) and one term when the maskers are symmetrically placed (i.e., binaural hearing is required for SRM). The Jones and Litovsky (2011) model includes similar descriptive terms but also has terms to deal with maskers that are not in the frontal azimuthal plane and terms for when maskers are both spatially separated from the target and co-located with the target. Both modeling attempts have constants that differ depending on several factors like the type of stimuli that are used (speech, noise, modulated noise, etc.). Neither the Bronkhorst (2000) nor the Jones and Litovsky (2011) model could account for the data of the present paper without having to arbitrarily change the values of the constants when either the condition changed or the number of masker sources changed. In general the models predicted more SRM than was obtained in the present experiments. Thus, in order for a descriptive model like those of Bronkhorst (2000) or Jones and Litovsky (2011) to better account for the data of the present paper both a function relating masking effects to increasing number of masker sources and masking effects related to increasing spatial separation of masker sources from the target source would have to be more compressive than either the predictions of the current Bronkhorst (2000) or Jones and Litovsky (2011) models. And, such a new model would be more useful than the existing ones if there were a way to account for differences in SRM due to informational and energetic masking. The current models can account for these differences only by changing the value of an arbitrary multiplicative constant. Developing a new descriptive model was not the intent of this mainly empirical paper.

Yost and Brown (2013) suggested a scheme that might be the basis of a process-like model for SRM. Their scheme was similar to other suggestions in the literature (e.g., Woodruff and Wang, 2010; Liu *et al.*, 2000). Yost and Brown (2013) surmised that localizing very similar sounds (i.e., independent noises) presented simultaneously from just two sources would be very difficult since the sound from the two sources would be mixed before they arrived at listeners' ears. They proposed a scheme in which the combined waveform was divided in a spectral/temporal matrix, where each cell in the matrix represented a brief slice of time and a narrow region of the spectrum. They used amplitude modulated noise bursts to show the many cells in a spectral/temporal matrix contained ITDs or ILDs that were similar to those existing in the spectral/temporal matrix representing the sound from only one source or the other. The ILDs and/or ITDs computed for many cells in the combined matrix were not similar to those of either source presenting sound in isolation. However, there were a lot of cells in the combined matrix that did represent the ITDs and/or ILDs of one source or the other. Yost and Brown (2013) suggested that these cells might be sufficient to allow for the localization of the sound sources, but if so, localization performance would be poor, as they showed. Yost and Brown (2013) explained

why additional information would be required to turn their scheme into a full-fledged model of multiple sound source localization.

In the Yost and Brown (2013) experiment, one of the two sounds could have been a target and the other sound a masker, and target detection/recognition in the presence of the spatially separated masker sources could have been evaluated. If so, the scheme they proposed could be used to imagine how the target and masker might be binaurally processed so that listeners could use the ITDs and/or ILDs (or some other binaural process, e.g., interaural coherence, see Faller and Merimaa, 2004; Dietz *et al.*, 2011) to localize each and improve the detection/recognition of the target. The data from Yost and Brown (2013) suggested that the temporal width of a cell in the spectral/temporal matrix could be as small as 2–5 ms and the spectral dimension on the order of an ERB. These cell dimensions could produce a large proportion of cells containing "reliable" estimates of the target and masker ITDs and/or ILDs as the basis for localizing each source (see Yost and Brown, 2013). However, as the number of sources increases the proportion of such "reliable" cells would decrease for fixed duration sounds. It would be essentially impossible for a scheme like that proposed by Yost and Brown (2013) and others (e.g., see Woodruff and Wang, 2010; Liu *et al.*, 2000) to produce a sufficient number of "reliable" cells for any more than three-four sources, assuming sounds of about a second in duration. Thus, it is likely that a model of SRM based on this or similar schemes would not be able to spatially segregate the location of more than three to four sources producing simultaneous sound. As discussed previously, this qualitatively agrees with the findings of this present study and those of Yost and Brown (2013), Santala and Pulkki (2011), and Kawashima and Sato (2015).

To the extent that SRM indicates an ability to effectively process sounds when competing sounds are spatially separated, the present data in concert with other data suggest that spatial separation may be useful only when there are five or fewer sound sources. This may not be a serious limitation of the use of SRM in the everyday world. First, it may be that we rarely deal with acoustic environments where there is a need to process five or more sources producing simultaneous sounds. Perhaps, most everyday listening scenarios involve attempts to process only a few sound sources. Second, the results of this study were limited to conditions in which the masker sources were symmetrically spaced. While such a spatial arrangement of sound sources can occur in the everyday world, such arrangements in which competing sound sources are only symmetrically spaced are probably rare. It is more likely that some competing sound sources would be asymmetrically spaced which would likely involve a role for head shadow effects. Thus, to more fully appreciate the influence of multiple sound sources on SRM in the everyday world, data like those collected in this study should probably be obtained with maskers that are both symmetrically and asymmetrically spaced relative to the target (e.g., similar to Bronkhorst and Plomp, 2000). Because the goal of this study was to better understand the role of binaural processing in SRM when there were multiple maskers, only symmetrically spaced maskers were used.

V. SUMMARY

When only binaural mechanisms are most likely used in processing target speech in the presence of either spatially separated speech or noise masker sources, spatial release from masking (SRM) decreases as the number of symmetrically spaced masker sources increases from two to six. There appears to be almost no SRM for six maskers. The results suggest that SRM decreases in a similar manner as the number of masker sources increases if masking is mainly energetic, informational, or a combination of energetic and informational masking. For the conditions of this study, speech-on-speech masking appears to be the sum of energetic masking (noise masker) and informational masking (target and masker spectrally non-overlapping). The results of the study appear to be consistent with other data and modeling attempts in the literature concerning sound source identification and localization of multiple spatially separated sound sources indicating that the spatial auditory scene probably contains fewer than five simultaneous sound sources.

ACKNOWLEDGMENTS

The research was supported by a grant from the Air Force Office of Scientific Research and a grant from the National Institute on Deafness and Other Communication Disorders both awarded to W.A.Y. The author is grateful to Dr. Michael Dorman, Dr. Yi Zhou, and Dr. Torben Pastore for comments on this research project. The assistance of Kathryn Pulling is also gratefully acknowledged.

[1]In some of the SRM literature the use of binaural processing is referred to as "binaural squelch," the opportunity to take advantage of head shadow is referred to as "head shadow" or the "better ear effect," and a third possibility for obtaining SRM is called "summation" in which a very small SRM might accrue based on both ears receiving redundant information (see Litovsky, 2012, for a further description of these terms). The use of these terms and concepts are often dependent on assuming that the two ears are independent receivers of acoustic information, an assumption that might not always hold (e.g., see Shub *et al.*, 2008a).

[2]Filtering and modulating noises based on the words was performed partially to allow for noise maskers to possibly be localized more accurately (especially when different noises were mixed together) than if they were just independently generated 750-ms noise bursts (see Yost and Brown, 2013). No tests of this assumption were carried out.

[3]Two-ERB wide filters were chosen based on pilot work. The aim was to find a filter as narrow as possible that would clearly differentiate the spectrum of one word from that of another word and still maintain a highly intelligible word. No formal tests were performed, but informal listening and measurement suggested that a two-ERB wide filter bank was a good compromise.

Arbogast, T. L., Mason, C. R., and Kidd, G., Jr. (**2002**). "The effect of spatial separation on informational and energetic masking of speech," J. Acoust. Soc. Am. **112**, 2086–2098.

Blauert, J. (**1997**). *Spatial Hearing* (MIT Press, Cambridge, MA), 494 pp.

Bronkhorst, A. W. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acta Acust. **86**, 117–128.

Bronkhorst, A. W., and Plomp, R. (**1992**). "Effect of multiple speech like maskers on binaural speech recognition in normal and impaired hearing," J. Acoust. Soc. Am. **92**, 3132–3139.

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.
Brungart, D. S., and Iyer, N. (**2012**). "Better-ear glimpsing efficiency with symmetrically-placed interfering talkers," J. Acoust. Soc. Am. **132**, 2545–2556.
Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**, 2527–2538.
Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975–979.
Cosentino, S., Marquardt, T., McAlpine, D., Culling, J. F., and Falk, T. H. (**2014**). "A model that predicts the binaural advantage to speech intelligibility from the mixed target and interferer signals," J. Acoust. Soc. Am. **135**, 796–807.
Dietz, M., Ewert, S. D., and Hohmann, V. (**2011**). "Auditory model based direction estimation of concurrent speakers from binaural signals," Speech Commun. **53**, 592–605.
Durlach, N. I. (**1963**). "Equalization and cancellation theory of binaural masking-level differences," J. Acoust. Soc. Am. **35**, 1206–1218.
Durlach, N. I., and Colburn, H. S. (**1978**). "Binaural phenomena," in *Handbook of Perception*, edited by E. C. Carterette and M. P. Friedman (Academic Press, New York), Vol. IV, pp. 365–463.
Faller, C., and Merimaa, J. (**2004**). "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," J. Acoust. Soc. Am. **116**, 3075–3089.
Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2004**). "Effect of number of masking talkers and auditory priming on informational masking in speech recognition," J. Acoust. Soc. Am. **115**, 2246–2256.
Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578–3588.
Green, D. M., and Yost, W. A. (**1975**). "Binaural analysis," in *Handbook of Sensory Physiology: Hearing*, edited by W. Keidel and W. D. Neff (Springer-Verlag, New York), Vol. 5, Chap. 11.
Hartmann, W. M., and Pumplin, J. (**1988**). "Noise power fluctuations and the masking of sine signals," J. Acoust. Soc. Am. **83**, 2277–2289.
Hirsh, I. J. (**1948**). "The influence of interaural phase on interaural summation and inhibition," J. Acoust. Soc. Am. **20**, 536–544.
Ihlefeld, A., and Shinn-Cunningham, B. G. (**2008**). "Spatial release from energetic and informational masking in a selective speech identification task," J. Acoust. Soc. Am. **123**, 4369–4379.
Jones, G. L., and Litovsky, R. Y. (**2011**). "A cocktail party model of spatial release from masking by both noise and speech interferers," J. Acoust. Soc. Am. **130**, 1463–1474.
Kawashima, T., and Sato, T. (**2015**). "Perceptual limits in a simulated 'Cocktail party,'" Atten. Percep. Psychol. **77**, 2108–2120.
Kidd, G., Jr., Mason, C. R., Richards, V. M., Gallun, F. J., and Durlach, N. I. (**2008**). "Informational masking," in *Auditory Perception of Sound Sources*, edited by W. A. Yost, A. N. Popper, and R. R. Fay (Springer, New York), Chap. 6.
Kidd, G., Jr., Mason, C. R., Swaminathan, S., Roverud, J. E., Kameron, K., Clayton, K. K., and Best, V. (**2016**). "Determining the energetic and informational components of speech-on-speech masking," J. Acoust. Soc. Am. **140**, 132–144.
Licklider, J. C. R. (**1948**). "The influence of interaural phase relations upon the masking of speech by white noise," J. Acoust. Soc. Am. **20**, 150–160.
Litovsky, R. Y. (**2012**). "Spatial release from masking," Acoust. Today **8**, 18–25.
Liu, C., Wheeler, B. C., O'Brien, W. D. O., Bilger, C., Lansing, C. R., and Feng, A. S. (**2000**). "Localization of multiple sources with two microphones," J. Acoust. Soc. Am. **108**, 1888–1905.
Lutfi, R., Gilbertson, L., Heo, I., Chang, A.-C., and Stamas, J. (**2013**). "The information-divergence hypothesis of informational masking," J. Acoust. Soc. Am. **134**, 2160–2170.
Santala, O., and Pulkki, V. (**2011**). "Directional perception of distributed sound sources," J. Acoust. Soc. Am. **129**, 1522–1530.
Shinn-Cunningham, B. G. (**2008**). "Object-based auditory and visual attention," Trends Cog. Sci. **12**, 182–186.
Shub, D. E., Durlach, N. I., and Colburn, H. S. (**2008a**). "Monaural level discrimination under dichotic conditions," J. Acoust. Soc. Am. **123**, 4421–4433.
Shub, D. E., Carr, S. P., Kong, Y., and Colburn, H. S. (**2008b**). "Discrimination and identification of azimuth using spectral shape," J. Acoust. Soc. Am. **124**, 3132–3141.
Srinivasan, N. K., Jakien, K. M., and Gallun, F. J. (**2016**). "Release from masking for small spatial separations: Effects of age and hearing loss," J. Acoust. Soc. Am. **140**, EL73–EL78.
Watson, C. S. (**2005**). "Some comments on informational masking," Acta Acust. **91**, 502–512.
Woodruff, J., and Wang, D. L. (**2010**). "Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization," IEEE Trans. Audiol. Speech Lang. Processing **18**, 1856–1866.
Yost, W. A. (**1997**). "The cocktail party effect: 40 years later," in *Localization and Spatial Hearing in Real and Virtual Environments*, edited by R. Gilkey and T. Anderson (Erlbaum Press, Mahwah, NJ), pp. 329–349.
Yost, W. A., and Brown, C. (**2013**). "Localizing the sources of two independent noises: Role of time varying amplitude differences," J. Acoust. Soc. Am. **133**, 2301–2313.
Yost, W. A., Zhong, X., and Najam, A. (**2015**). "Rotating sound sources and listeners: Sound source localization is a multisensory/cognitive process," J. Acoust. Soc. Am. **137**, 2200–2201.
Zurek, P. M. (**1993**). "Binaural advantages and directional effects in speech intelligibility," in *Acoustical Factors affecting Hearing Aid Performance*, edited by G. A. Studebaker and I. Hochberg (Allyn and Bacon, Boston, MA), pp. 255–276.